

# Surrogate regret bounds for generalized classification performance metrics\*

Wojciech Kotłowski

Poznan University of Technology, Poland

Krzysztof Dembczyński

Poznan University of Technology, Poland

October 10, 2016

## Abstract

We consider optimization of generalized performance metrics for binary classification by means of surrogate losses. We focus on a class of metrics, which are linear-fractional functions of the false positive and false negative rates (examples of which include  $F_\beta$ -measure, Jaccard similarity coefficient, AM measure, and many others). Our analysis concerns the following two-step procedure. First, a real-valued function  $f$  is learned by minimizing a surrogate loss for binary classification on the training sample. It is assumed that the surrogate loss is a strongly proper composite loss function (examples of which include logistic loss, squared-error loss, exponential loss, etc.). Then, given  $f$ , a threshold  $\hat{\theta}$  is tuned on a separate validation sample, by direct optimization of the target performance metric. We show that the regret of the resulting classifier (obtained from thresholding  $f$  on  $\hat{\theta}$ ) measured with respect to the target metric is upperbounded by the regret of  $f$  measured with respect to the surrogate loss. We also extend our results to cover multilabel classification and provide regret bounds for micro- and macro-averaging measures. Our findings are further analyzed in a computational study on both synthetic and real data sets.

## 1 Introduction

In binary classification, misclassification error is not necessarily an adequate evaluation metric, and one often resorts to more complex metrics, better suited for the problem. For instance, when classes are imbalanced,  $F_\beta$ -measure (Lewis, 1995; Jansche, 2005; Nan et al, 2012) and AM measure (balanced error rate) (Menon et al, 2013) are frequently used. Optimizing such generalized performance metrics poses computational and statistical challenges, as they cannot be decomposed into losses on individual observations.

In this paper, we consider optimization of generalized performance metrics by means of surrogate losses. We restrict our attention to a family of performance metrics which are ratios of linear functions of false positives (FP) and false negatives (FN). Such functions are called linear-fractional, and include the aforementioned  $F_\beta$  and AM measures, as well as Jaccard similarity coefficient, weighted accuracy, and many others (Koyejo et al, 2014, 2015). We focus on the most popular approach to optimizing generalized performance metrics in practice, based on the following two-step procedure. First, a real-valued function  $f$  is learned by minimizing a surrogate loss for binary classification on the training sample. Then, given  $f$ , a threshold  $\hat{\theta}$  is tuned on a separate validation sample, by direct optimization of the target performance measure with respect to a classifier obtained from  $f$  by thresholding at  $\hat{\theta}$ , classifying all observations with value of  $f$  above the threshold as positive class, and all observations below the threshold as negative class. This approach can be motivated by the asymptotic analysis: minimization of appropriate surrogate loss results in estimation of conditional (“posterior”) class probabilities, and many performance metrics are maximized by a classifier which predicts by thresholding on the scale of conditional probabilities (Nan et al, 2012; Zhao et al, 2013;

---

\*W. Kotłowski has been supported by the Polish National Science Centre under grant no. 2013/11/D/ST6/03050. K. Dembczyński has been supported by the Polish National Science Centre under grant no. 2013/09/D/ST6/03917.

Koyejo et al, 2014). However, it is unclear what can be said about the behavior of this procedure on finite samples.

In this work, we are interested in theoretical analysis and justification of this approach for any sample size, and for any, not necessarily perfect, classification function. To this end, we use the notion of *regret* with respect to some evaluation metric, which is a difference between the performance of a given classifier and the performance of the optimal classifier with respect to this metric. We show that the regret of the resulting classifier (obtained from thresholding  $f$  on  $\hat{\theta}$ ) measured with respect to the target metric is upperbounded by the regret of  $f$  measured with respect to the surrogate loss. Our result holds for any surrogate loss function, which is *strongly proper composite* (Agarwal, 2014), examples of which include logistic loss, squared-error loss, exponential loss, etc. Interestingly, the proof of our result goes by an intermediate bound of the regret with respect to the target measure by a cost-sensitive classification regret. As a byproduct, we get a bound on the cost-sensitive classification regret by a surrogate regret of a real-valued function which holds *simultaneously* for *all* misclassification costs: the misclassification costs only influence the threshold, but not: the function, the surrogate loss, or the regret bound.

We further extend our results to cover multilabel classification, in which the goal is to simultaneously predict multiple labels for each object. We consider two methods of generalizing binary classification performance metrics to the multilabel setting: the macro-averaging and the micro-averaging (Manning et al, 2008; Parambath et al, 2014; Koyejo et al, 2015). The macro-averaging is based on first computing the performance metric separately for each label, and then averaging the metrics over the labels. In the micro-averaging, the false positives and false negatives for each label are first averaged over the labels, and then the performance metric is calculated on these averaged quantities. We show that our regret bounds hold for both macro- and micro-averaging measures. Interestingly, for micro averaging, only a single threshold needs to be tuned and is shared among all labels.

Our finding is further analyzed in a computational study on both synthetic and real data sets. We compare the performance of the algorithm when used with two types of surrogate losses: the logistic loss (which is strongly proper) and the hinge loss (which is not a proper loss). On synthetic data sets, we analyze the behavior of the algorithm for discrete feature distribution (where nonparametric classifiers are used), and for continuous feature distribution (where linear classifiers are used). Next, we look at the performance of the algorithm on the real-life benchmark data sets, both for binary and multilabel classification.

We note that the goal of this paper is not to propose a new learning algorithm, but rather to provide a deeper statistical understanding of an existing method. The two-step procedure analyzed here (also known as the plug-in method in the case when the outcomes of the function have a probabilistic interpretation), is commonly used in the binary classification with generalized performance metrics, but this is exactly the reason why we think it is important to study this method in more depth from a theoretical point of view.

## 1.1 Related work

In machine learning, numerous attempts to optimize generalized performance metrics have been proposed. They can be divided into two general categories. The structured loss approaches (Musicant et al, 2003; Tsochantaridis et al, 2005; Petterson and Caetano, 2011, 2010) rely on incorporating the performance metric into the training process, thus requiring specialized learning algorithms to optimize non-standard objectives. On the other hand, the plug-in approaches, which are very closely related to the topic of this work, are based on obtaining reliable class conditional probability estimates by employing standard algorithms minimizing some surrogate loss for binary classification (such as logistic loss used in logistic regression, exponential loss used in boosting, etc.), and then plugging these estimates into the functional form of the optimal prediction rule for a given performance metric (Jansche, 2007; Nan et al, 2012; Dembczyński et al, 2013; Waegeman et al, 2013; Narasimhan et al, 2014, 2015; Koyejo et al, 2014, 2015).

Existing theoretical work on generalized performance metrics is mainly concerned with *statistical consistency* also known as *calibration*, which determines whether convergence to the minimizer of a surrogate loss implies convergence to the minimizer of the task performance measure as the sample size goes to infinity (Dembczyński et al, 2010; Nan et al, 2012; Gao and Zhou, 2013; Zhao et al, 2013; Narasimhan et al, 2014;

Koyejo et al, 2014, 2015). Here we give a stronger result which bounds the regret with respect to the performance metric by the regret with respect to the surrogate loss. Our result is valid for all finite sample sizes and informs about the rates of convergence.

We also note that two distinct frameworks are used to study the statistical consistency of classifiers with respect to performance metrics: Decision Theoretic Analysis (DTA), which assumes a test set of a fixed size, and Empirical Utility Maximization (EUM), in which the metric is defined by means of population quantities (Nan et al, 2012). In this context, our work falls into the EUM framework.

Parambath et al (2014) presented an alternative approach to maximizing linear-fractional metrics by learning a sequence of binary classification problems with varying misclassification costs. While we were inspired by their theoretical analysis, their approach is, however, more complicated than the two-step approach analyzed here, which requires solving an ordinary binary classification problem only once. Moreover, as part of our proof, we show that by minimizing a strongly proper composite loss, we are *implicitly* minimizing cost-sensitive classification error for any misclassification costs without any overhead. Hence, the costs need not be known during learning, and can only be determined later on a separate validation sample by optimizing the threshold. Narasimhan et al (2015) developed a general framework for designing provably consistent algorithms for complex multiclass performance measures. They relate the regret with respect to the target metric to the conditional probability estimation error measured in terms of  $L_1$ -metric. Their algorithms rely on using accurate class conditional probability estimates and multiple solving cost-sensitive multiclass classification problems.

The generalized performance metrics for binary classification are employed in the multilabel setting by means of one of the three averaging schemes (Waegeman et al, 2013; Parambath et al, 2014; Koyejo et al, 2015): instance-averaging (averaging errors over the labels, averaging metric over the examples), macro-averaging (averaging errors over the examples, averaging metric over the labels), and micro-averaging (averaging errors over the examples and the labels). Koyejo et al (2015) characterize the optimal classifiers for multilabel metrics and prove the consistency of the plug-in method. Our regret bounds for multilabel classification can be seen as a follow up on their work.

## 1.2 Outline

The paper is organized as follows. In Section 2 we introduce basic concepts, definitions and notation. The main result is presented in Section 3 and proved in Section 4. Section 5 extends our results to the multilabel setting. The theoretical contribution of the paper is complemented by computational experiments in Section 6, prior to concluding with a summary in Section 7.

## 2 Problem setting

### 2.1 Binary classification

In binary classification, the goal is, given an input (feature vector)  $x \in X$ , to accurately predict the output (label)  $y \in \{-1, 1\}$ . We assume input-output pairs  $(x, y)$  are generated i.i.d. according to  $\Pr(x, y)$ . A *classifier* is a mapping  $h: X \rightarrow \{-1, 1\}$ . Given  $h$ , we define the following four quantities:

$$\begin{aligned} \text{TP}(h) &= \Pr(h(x) = 1 \wedge y = 1), \\ \text{FP}(h) &= \Pr(h(x) = 1 \wedge y = -1), \\ \text{TN}(h) &= \Pr(h(x) = -1 \wedge y = -1), \\ \text{FN}(h) &= \Pr(h(x) = -1 \wedge y = 1), \end{aligned}$$

which are known as *true positives*, *false positives*, *true negatives* and *false negatives*, respectively. We also denote  $\Pr(y = 1)$  by  $P$ . Note that for any  $h$ ,  $\text{FP}(h) + \text{TN}(h) = \Pr(y = -1) = 1 - P$  and  $\text{TP}(h) + \text{FN}(h) = P$ , so out of the four quantities above, only two are independent. In this paper, we use the convention to parameterize all metrics by means of  $\text{FP}(h)$  and  $\text{FN}(h)$ .

metric	expression
Accuracy	$\text{Acc} = 1 - \text{FN} - \text{FP}$
$F_\beta$ -measure	$F_\beta = \frac{(1+\beta^2)(P-\text{FN})}{(1+\beta^2)P-\text{FN}+\text{FP}}$
Jaccard similarity	$J = \frac{P-\text{FN}}{P+\text{FP}}$
AM measure	$\text{AM} = \frac{2P(1-P)-P\text{FP}-(1-P)\text{FN}}{2P(1-P)}$
Weighted accuracy	$\text{WA} = \frac{w_1(1-P)+w_2P-w_1\text{FP}-w_2\text{FN}}{w_1(1-P)+w_2P}$

Table 1: Some popular linear-fractional performance metrics expressed as functions of FN and FP. See (Koyejo et al, 2014) for a more detailed description.

We call a two-argument function  $\Psi = \Psi(\text{FP}, \text{FN})$  a (*generalized*) *classification performance metric*. Given a classifier  $h$ , we define  $\Psi(h) = \Psi(\text{FP}(h), \text{FN}(h))$ . Throughout the paper we assume that  $\Psi(\text{FP}, \text{FN})$  is *linear-fractional*, i.e., is a ratio of linear functions:

$$\Psi(\text{FP}, \text{FN}) = \frac{a_0 + a_1\text{FP} + a_2\text{FN}}{b_0 + b_1\text{FP} + b_2\text{FN}}, \quad (1)$$

where we allow coefficients  $a_i, b_i$  to depend on the distribution  $\Pr(x, y)$ . Note, that our convention to parameterize the metric by means of  $(\text{FP}, \text{FN})$  does not affect definition (1), because  $\Psi$  can be reparameterized to  $(\text{FP}, \text{TN})$ ,  $(\text{TP}, \text{FN})$ , or  $(\text{TP}, \text{TN})$ , and will remain linear-fractional in all these parameterizations. We also assume  $\Psi(\text{FP}, \text{FN})$  is non-increasing in FP and FN, a property that is inherently possessed by virtually all performance measures used in practice. Table 1 lists some popular examples of linear-fractional performance metrics.

Let  $h_\Psi^*$  be the maximizer of  $\Psi(h)$  over all classifiers:

$$h_\Psi^* = \underset{h: X \rightarrow \{-1, 1\}}{\text{argmax}} \Psi(h)$$

(if argmax is not unique, we take  $h_\Psi^*$  to be any maximizer of  $\Psi$ ). Given any classifier  $h$ , we define its  $\Psi$ -regret as:

$$\text{Reg}_\Psi(h) = \Psi(h_\Psi^*) - \Psi(h).$$

The  $\Psi$ -regret is nonnegative from the definition, and quantifies the suboptimality of  $h$ , i.e., how much worse is  $h$  comparing to the optimal  $h_\Psi^*$ .

## 2.2 Strongly proper composite losses

Here we briefly outline the theory of strongly proper composite loss functions. See (Agarwal, 2014) for a more detailed description.

Define a *binary class probability estimation (CPE) loss function* (Reid and Williamson, 2010, 2011) as a function  $c: \{-1, 1\} \times [0, 1] \rightarrow \mathbb{R}_+$ , where  $c(y, \hat{\eta})$  assigns penalty to prediction  $\hat{\eta}$ , when the observed label is  $y$ . Define the *conditional c-risk* as:<sup>1</sup>

$$\text{risk}_c(\eta, \hat{\eta}) = \eta c(1, \hat{\eta}) + (1 - \eta) c(-1, \hat{\eta}),$$

the expected loss of prediction  $\hat{\eta}$  when the label is drawn from a distribution with  $\Pr(y = 1) = \eta$ . We say CPE loss is *proper* if for any  $\eta \in [0, 1]$ ,  $\eta \in \text{argmin}_{\hat{\eta} \in [0, 1]} \text{risk}_c(\eta, \hat{\eta})$ . In other words, proper losses are

<sup>1</sup>Throughout the paper, we follow the convention that all conditional quantities are lowercase (regret, risk), while all unconditional quantities are uppercase (Regret, Risk).

minimized by taking the true class probability distribution as a prediction; hence  $\hat{\eta}$  can be interpreted as probability estimate of  $\eta$ . Define the *conditional c-regret* as:

$$\begin{aligned}\text{reg}_c(\eta, \hat{\eta}) &= \text{risk}_c(\eta, \hat{\eta}) - \inf_{\hat{\eta}'} \text{risk}_c(\eta, \hat{\eta}') \\ &= \text{risk}_c(\eta, \hat{\eta}) - \text{risk}_c(\eta, \eta),\end{aligned}$$

the difference between the conditional  $c$ -risk of  $\hat{\eta}$  and the optimal  $c$ -risk. We say a CPE loss  $c$  is  $\lambda$ -strongly proper if for any  $\eta, \hat{\eta}$ :

$$\text{reg}_c(\eta, \hat{\eta}) \geq \frac{\lambda}{2}(\eta - \hat{\eta})^2,$$

i.e. the conditional  $c$ -regret is everywhere lowerbounded by a squared difference of its arguments. It can be shown (Agarwal, 2014) that under mild regularity assumption a proper CPE loss  $c$  is  $\lambda$ -strongly proper if and only if the function  $H_c(\eta) := \text{risk}_c(\eta, \eta)$  is  $\lambda$ -strongly concave. This fact lets us easily verify whether a given loss function is  $\lambda$ -strongly proper.

It is often more convenient to reparameterize the loss function from  $\hat{\eta} \in [0, 1]$  to a real-valued  $f \in \mathbb{R}$  through a strictly increasing (and therefore invertible) *link function*  $\psi: [0, 1] \rightarrow \mathbb{R}$ :

$$\ell(y, f) = c(y, \psi^{-1}(f)).$$

If  $c$  is  $\lambda$ -strongly proper, we call function  $\ell: \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R}_+$   *$\lambda$ -strongly proper composite loss function*. The notions of conditional  $\ell$ -risk  $\text{risk}_\ell(\eta, f)$  and conditional  $\ell$ -regret  $\text{reg}_\ell(\eta, f)$  extend naturally to the case of composite losses:

$$\begin{aligned}\text{risk}_\ell(\eta, f) &= \eta \ell(1, f) + (1 - \eta) \ell(-1, f) \\ \text{reg}_\ell(\eta, f) &= \text{risk}_\ell(\eta, f) - \inf_{f' \in \mathbb{R}} \text{risk}_\ell(\eta, f') \\ &= \text{risk}_\ell(\eta, f) - \text{risk}_\ell(\eta, \psi(\eta)).\end{aligned}$$

and the strong properness of underlying CPE loss implies:

$$\text{reg}_\ell(\eta, f) \geq \frac{\lambda}{2} \left( \eta - \psi^{-1}(f) \right)^2 \quad (2)$$

As an example, consider a *logarithmic scoring rule*:

$$c(y, \hat{\eta}) = -\mathbb{I}[y = 1] \log \hat{\eta} - \mathbb{I}[y = -1] \log(1 - \hat{\eta}),$$

where  $\mathbb{I}[Q]$  is the indicator function, equal to 1 if  $Q$  holds, and to 0 otherwise. Its conditional risk is given by:

$$\text{risk}_c(\eta, \hat{\eta}) = -\eta \log \hat{\eta} - (1 - \eta) \log(1 - \hat{\eta}),$$

the *cross-entropy* between  $\eta$  and  $\hat{\eta}$ . The conditional  $c$ -regret is the binary *Kullback-Leibler divergence* between  $\eta$  and  $\hat{\eta}$ :

$$\text{reg}_c(\eta, \hat{\eta}) = \eta \log \frac{\eta}{\hat{\eta}} + (1 - \eta) \log \frac{1 - \eta}{1 - \hat{\eta}}.$$

Note that since  $H(\eta) = \text{risk}_c(\eta, \eta)$  is the binary entropy function, and  $\left| \frac{d^2 H}{d\eta^2} \right| = \frac{1}{\eta(1-\eta)} \geq 4$ ,  $c$  is 4-strongly proper loss. Using the *logit* link function  $\psi(\hat{\eta}) = \log \frac{\hat{\eta}}{1-\hat{\eta}}$ , we end up with the logistic loss function:

$$\ell(y, f) = \log \left( 1 + e^{-yf} \right),$$

which is 4-strongly proper composite from the definition.

Table 2 presents some of the commonly used losses which are strongly proper composite. Note that the *hinge loss*  $\ell(y, f) = (1 - yf)_+$ , used, e.g., in support vector machines (Hastie et al, 2009), is *not* strongly proper composite (even not proper composite).

loss function	squared-error	logistic	exponential
$\ell(y, f)$	$(y - f)^2$	$\log(1 + e^{-fy})$	$e^{-yf}$
$c(1, \hat{\eta})$	$4(1 - \hat{\eta})^2$	$-\log \hat{\eta}$	$\sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}}$
$c(-1, \hat{\eta})$	$4\hat{\eta}^2$	$-\log(1 - \hat{\eta})$	$\sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}}$
$\psi(\hat{\eta})$	$2\hat{\eta} - 1$	$\log \frac{\hat{\eta}}{1-\hat{\eta}}$	$\frac{1}{2} \log \frac{\hat{\eta}}{1-\hat{\eta}}$
$\lambda$	8	4	4

Table 2: Three popular strongly proper composite losses: squared-error, logistic and exponential losses. Shown are the formula  $\ell(y, f)$ , the underlying CPE loss  $c(y, \hat{\eta})$  with the link function  $\psi(\hat{\eta})$ , as well as the strong properness constant  $\lambda$ . See (Agarwal, 2014) for more details and examples.

### 3 Main result

Given a real-valued function  $f: X \rightarrow \mathbb{R}$ , and a  $\lambda$ -strongly proper composite loss  $\ell(y, f)$ , define the  $\ell$ -risk of  $f$  as the expected loss of  $f(x)$  with respect to the data distribution:

$$\begin{aligned} \text{Risk}_\ell(f) &= \mathbb{E}_{(x,y)} [\ell(y, f(x))] \\ &= \mathbb{E}_x [\text{risk}_\ell(\eta(x), f(x))] , \end{aligned}$$

where  $\eta(x) = \Pr(y = 1|x)$ . Let  $f_\ell^*$  be the minimizer  $\text{Risk}_\ell(f)$  over all functions,  $f_\ell^* = \text{argmin}_f \text{Risk}_\ell(f)$ . Since  $\ell$  is proper composite:

$$f_\ell^*(x) = \psi(\eta(x)).$$

Define the  $\ell$ -regret of  $f$  as:

$$\begin{aligned} \text{Reg}_\ell(f) &= \text{Risk}_\ell(f) - \text{Risk}_\ell(f_\ell^*) \\ &= \mathbb{E}_x [\text{risk}_\ell(\eta(x), f(x)) - \text{risk}_\ell(\eta(x), f_\ell^*(x))] . \end{aligned}$$

Any real-valued function  $f: X \rightarrow \mathbb{R}$  can be turned into a classifier  $h_{f,\theta}: X \rightarrow \{-1, 1\}$ , by thresholding at some value  $\theta$ :

$$h_{f,\theta}(x) = \text{sgn}(f(x) - \theta).$$

The purpose of this paper is to address the following problem: given a function  $f$  with  $\ell$ -regret  $\text{Reg}_\ell(f)$ , and a threshold  $\theta$ , what can we say about  $\Psi$ -regret of  $h_{f,\theta}$ ? For instance, can we bound  $\text{Reg}_\Psi(h_{f,\theta})$  in terms of  $\text{Reg}_\ell(f)$ ? We give a positive answer to this question, which is based on the following regret bound:

**Lemma 1.** *Let  $\Psi(\text{FP}, \text{FN})$  be a linear-fractional function of the form (1), which is non-increasing in FP and FN. Assume that there exists  $\gamma > 0$ , such that for any classifier  $h: X \rightarrow \{-1, 1\}$ :*

$$b_0 + b_1 \text{FP}(h) + b_2 \text{FN}(h) \geq \gamma,$$

*i.e. the denominator of  $\Psi$  is positive and bounded away from zero. Let  $\ell$  be a  $\lambda$ -strongly proper composite loss function. Then, there exists a threshold  $\theta^*$ , such that for any real-valued function  $f: X \rightarrow \mathbb{R}$ ,*

$$\text{Reg}_\Psi(h_{f,\theta^*}) \leq C \sqrt{\frac{2}{\lambda}} \sqrt{\text{Reg}_\ell(f)},$$

*where  $C = \frac{1}{\gamma} (\Psi(h_\Psi^*)(b_1 + b_2) - (a_1 + a_2)) > 0$ .*

metric	$\gamma$	$C$
Accuracy	1	2
$F_\beta$ -measure	$\beta^2 P$	$\frac{1+\beta^2}{\beta^2 P}$
Jaccard similarity	$P$	$\frac{J^*+1}{P}$
AM measure	$2P(1-P)$	$\frac{1}{2P(1-P)}$
Weighted accuracy	$w_1 P + w_2(1-P)$	$\frac{w_1+w_2}{w_1 P + w_2(1-P)}$

Table 3: Constants which appear in the bound of Lemma 1 for several performance metrics.

The proof is quite long and hence is postponed to Section 4. Interestingly, the proof goes by an intermediate bound of the  $\Psi$ -regret by a cost-sensitive classification regret. We note that the bound in Lemma 1 is in general unimprovable, in the sense that it is easy to find  $f$ ,  $\Psi$ ,  $\ell$ , and distribution  $\Pr(x, y)$ , for which the bound holds with equality (see proof for details). We split the constant in front of the bound into  $C$  and  $\lambda$ , because  $C$  depends only on  $\Psi$ , while  $\lambda$  depends only on  $\ell$ . Table 3 lists these constants for some popular metrics. We note that constant  $\gamma$  (lower bound on the denominator of  $\Psi$ ) will be *distribution-dependent* in general (as it can depend on  $P = \Pr(y = 1)$ ) and may not have a uniform lower bound which holds for all distributions.

Lemma 1 has the following interpretation. If we are able to find a function  $f$  with small  $\ell$ -regret, we are guaranteed that there exists a threshold  $\theta^*$  such that  $h_{f, \theta^*}$  has small  $\Psi$ -regret. Note that the same threshold  $\theta^*$  will work for any  $f$ , and the right hand side of the bound is *independent* of  $\theta^*$ . Hence, to minimize the right hand side we only need to minimize  $\ell$ -regret, and we can deal with the threshold afterwards.

Lemma 1 also reveals the form of the optimal classifier  $h_\Psi^*$ : take  $f = f_\ell^*$  in the lemma and note that  $\text{Reg}_\ell(f_\ell^*) = 0$ , so that  $\text{Reg}_\Psi(h_{f_\ell^*, \theta^*}) = 0$ , which means that  $h_{f_\ell^*, \theta^*}$  is the minimizer of  $\Psi$ :

$$h_\Psi^*(x) = \text{sgn}(f_\ell^*(x) - \theta^*) = \text{sgn}(\eta(x) - \psi^{-1}(\theta^*)),$$

where the second equality is due to  $f_\ell^* = \psi(\eta)$  and strict monotonicity of  $\psi$ . Hence,  $h_\Psi^*$  is a threshold function on  $\eta$ . The proof of Lemma 1 (see Section 4) actually specifies the exact value of the threshold  $\theta^*$ :

$$\psi^{-1}(\theta^*) = \frac{\Psi(h_\Psi^*)b_1 - a_1}{\Psi(h_\Psi^*)(b_1 + b_2) - (a_1 + a_2)}, \quad (3)$$

which is in agreement with the result obtained by Koyejo et al (2014).<sup>2</sup>

To make Lemma 1 easier to grasp, consider a special case when the performance metric  $\Psi(\text{FP}, \text{FN}) = 1 - \text{FP} - \text{FN}$  is the classification accuracy. In this case, (3) gives  $\Psi^{-1}(\theta^*) = 1/2$ . Hence, we obtained the well-known result that the classifier maximizing the accuracy is a threshold function on  $\eta$  at  $1/2$ . Then, Lemma 1 states that given a real-valued  $f$ , we should take a classifier  $h_{f, \theta^*}$  which thresholds  $f$  at  $\theta^* = \psi(1/2)$ . Using Table 2, one can easily verify that  $\theta^* = 0$  for logistic, squared-error and exponential loss. This agrees with the common approach of thresholding the real-valued classifiers trained by minimizing these losses at 0 to obtain the label prediction. The bounds from the lemma are in this case identical (up to a multiplicative constant) to the bounds obtained by Bartlett et al (2006).

Unfortunately, for more complicated performance metrics, the optimal threshold  $\theta^*$  is unknown, as (3) contains unknown quantity  $\Psi(h_\Psi^*)$ , the value of the metric at optimum. The solution in this case is to, given  $f$ , directly search for a threshold which maximizes  $\Psi(h_{f, \theta})$ . This is the main result of the paper:

<sup>2</sup>To prove (3), Koyejo et al (2014) require an absolute continuity assumption on the marginal distribution over instances with respect to some dominating measure, so as to guarantee the existence of an appropriate density. Our analysis shows that the existence of a density is not required.



**Theorem 2.** *Given a real-valued function  $f$ , let  $\theta_f^* = \operatorname{argmax}_\theta \Psi(h_{f,\theta})$ . Then, under the assumptions and notation from Lemma 1:*

$$\operatorname{Reg}_\Psi(h_{f,\theta_f^*}) \leq C \sqrt{\frac{2}{\lambda}} \sqrt{\operatorname{Reg}_\ell(f)}.$$

*Proof.* The result follows immediately from Lemma 1: Solving  $\max_\theta \Psi(h_{f,\theta})$  is equivalent to solving  $\min_\theta \operatorname{Reg}_\Psi(h_{f,\theta})$ , and  $\min_\theta \operatorname{Reg}_\Psi(h_{f,\theta}) \leq \operatorname{Reg}_\Psi(h_{f,\theta^*})$ , where  $\theta^*$  is the threshold given by Lemma 1.  $\square$

Theorem 2 motivates the following procedure for maximization of  $\Psi$ :

1. Find  $f$  with small  $\ell$ -regret, e.g. by using a learning algorithm minimizing  $\ell$ -risk on the training sample.
2. Given  $f$ , solve  $\theta_f^* = \operatorname{argmax}_\theta \Psi(h_{f,\theta})$ .

Theorem 2 states that the  $\Psi$ -regret of the classifier obtained by this procedure is upperbounded by the  $\ell$ -regret of the underlying real-valued function.

We now discuss how to approach step 2 of the procedure in practice. In principle, this step requires maximizing  $\Psi$  defined through FP and FN, which are expectations over an unknown distribution  $\Pr(x, y)$ . However, it is sufficient to optimize  $\theta$  on the empirical counterpart of  $\Psi$  calculated on a separate validation sample. Let  $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$  be the validation set of size  $n$ . Define:

$$\widehat{\text{FP}}(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[h(x_i) = 1, y_i = -1], \quad \widehat{\text{FN}}(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[h(x_i) = -1, y_i = 1],$$

the empirical counterparts of FP and FN, and let  $\widehat{\Psi}(h) = \Psi(\widehat{\text{FP}}(h), \widehat{\text{FN}}(h))$  be the empirical counterpart of the performance metric  $\Psi$ . We now replace step 2 by:

Given  $f$  and validation sample  $\mathcal{T}$ , solve  $\widehat{\theta}_f = \operatorname{argmax}_\theta \widehat{\Psi}(h_{f,\theta})$ .

In Theorem 3 below, we show that:

$$\operatorname{Reg}_\Psi(h_{f,\widehat{\theta}_f}) - \operatorname{Reg}_\Psi(h_{f,\theta_f^*}) = O\left(\sqrt{\frac{\log n}{n}}\right),$$

so that tuning the threshold on the validation sample of size  $n$  (which results in  $\widehat{\theta}_f$ ) instead of on the population level (which results in  $\theta_f^*$ ) will cost at most  $O\left(\sqrt{\frac{\log n}{n}}\right)$  additional regret. The main idea of the proof is that finding the optimal threshold comes down to optimizing within a class of  $\{-1, 1\}$ -valued threshold functions, which has small Vapnik-Chervonenkis dimension. This, together with the fact that under assumptions from Lemma 1,  $\Psi$  is stable with respect to its arguments, implies that  $\Psi(h_{f,\widehat{\theta}_f})$  is close to  $\Psi(h_{f,\theta_f^*})$ .

**Theorem 3.** *Let the assumptions from Lemma 1 hold, and let:*

$$D_1 = \sup_{(\text{FP}, \text{FN})} |b_1 \Psi(\text{FP}, \text{FN}) - a_1|, \quad D_2 = \sup_{(\text{FP}, \text{FN})} |b_2 \Psi(\text{FP}, \text{FN}) - a_2|,$$

and  $D = \max\{D_1, D_2\}$ . Given a real-valued function  $f$ , and a validation set  $\mathcal{T}$  of size  $n$  generated i.i.d. from  $P(x, y)$ , let  $\widehat{\theta}_f = \operatorname{argmax}_\theta \widehat{\Psi}(h_{f,\theta})$  be the threshold maximizing the empirical counterpart of  $\Psi$  evaluated on  $\mathcal{T}$ . Then, with probability  $1 - \delta$  (over the random choice of  $\mathcal{T}$ ):

$$\operatorname{Reg}_\Psi(h_{f,\widehat{\theta}_f}) \leq C \sqrt{\frac{2}{\lambda}} \sqrt{\operatorname{Reg}_\ell(f)} + \frac{16D}{\gamma} \sqrt{\frac{4(1 + \log n) + 2 \log \frac{16}{\delta}}{n}}.$$



*Proof.* For any FP and FN, we have:

$$\begin{aligned} \left| \frac{\partial \Psi(\text{FP}, \text{FN})}{\partial \text{FP}} \right| &= \frac{|a_1(b_0 + b_1 \text{FP} + b_2 \text{FN}) - b_1(a_0 + a_1 \text{FP} + a_2 \text{FN})|}{(b_0 + b_1 \text{FP} + b_2 \text{FN})^2} \\ &= \frac{|b_1 \Psi(\text{FP}, \text{FN}) - a_1|}{b_0 + b_1 \text{FP} + b_2 \text{FN}} \leq \frac{|b_1 \Psi(\text{FP}, \text{FN}) - a_1|}{\gamma} \leq \frac{D}{\gamma}, \end{aligned}$$

and similarly,

$$\left| \frac{\partial \Psi(\text{FP}, \text{FN})}{\partial \text{FN}} \right| = \frac{|b_2 \Psi(\text{FP}, \text{FN}) - a_2|}{b_0 + b_1 \text{FP} + b_2 \text{FN}} \leq \frac{D}{\gamma}.$$

For any (FP, FN) and (FP', FN'), Taylor-expanding  $\Psi(\text{FP}, \text{FN})$  around (FP', FN') up to the first order and using the bounds above gives:

$$\Psi(\text{FP}, \text{FN}) \leq \Psi(\text{FP}', \text{FN}') + \frac{D}{\gamma} (|\text{FP} - \text{FP}'| + |\text{FN} - \text{FN}'|). \quad (4)$$

Now, we have:

$$\begin{aligned} \text{Reg}_\Psi(h_{f, \hat{\theta}_f}) &= \text{Reg}_\Psi(h_{f, \theta_f^*}) + \Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f}) \\ &\leq C \sqrt{\frac{2}{\lambda}} \sqrt{\text{Reg}_\ell(f)} + \Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f}), \end{aligned}$$

where we used Theorem 2. Thus, it amounts to bound  $\Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f})$ . From the definition of  $\hat{\theta}_f$ ,  $\hat{\Psi}(h_{f, \hat{\theta}_f}) \geq \hat{\Psi}(h_{f, \theta_f^*})$ , hence:

$$\begin{aligned} \Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f}) &\leq \Psi(h_{f, \theta_f^*}) - \hat{\Psi}(h_{f, \theta_f^*}) + \hat{\Psi}(h_{f, \hat{\theta}_f}) - \Psi(h_{f, \hat{\theta}_f}) \\ &\leq 2 \sup_{\theta} |\Psi(h_{f, \theta}) - \hat{\Psi}(h_{f, \theta})| \\ &= 2 \sup_{\theta} |\Psi(\text{FP}(h_{f, \theta}), \text{FN}(h_{f, \theta})) - \Psi(\widehat{\text{FP}}(h_{f, \theta}), \widehat{\text{FN}}(h_{f, \theta}))|, \end{aligned}$$

where we used the definition of  $\hat{\Psi}$ . Using (4),

$$\Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f}) \leq \frac{2D}{\gamma} \left( \sup_{\theta} |\text{FP}(h_{f, \theta}) - \widehat{\text{FP}}(h_{f, \theta})| + \sup_{\theta} |\text{FN}(h_{f, \theta}) - \widehat{\text{FN}}(h_{f, \theta})| \right).$$

Note that the suprema above are on the deviation of empirical mean from the expectation over the class of threshold functions, which has Vapnik-Chervonenkis dimension equal to 2. Using standard argument from Vapnik-Chervonenkis theory (see, e.g., [Devroye et al, 1996](#)), with probability  $1 - \frac{\delta}{2}$  over the random choice of  $\mathcal{T}$ :

$$\sup_{\theta} |\text{FP}(h_{f, \theta}) - \widehat{\text{FP}}(h_{f, \theta})| \leq 4 \sqrt{\frac{4(1 + \log n) + 2 \log \frac{16}{\delta}}{n}},$$

and similarly for the second supremum. Thus, with probability  $1 - \delta$ ,

$$\Psi(h_{f, \theta_f^*}) - \Psi(h_{f, \hat{\theta}_f}) \leq \frac{16D}{\gamma} \sqrt{\frac{4(1 + \log n) + 2 \log \frac{16}{\delta}}{n}},$$

which finishes the proof.  $\square$

We note that, contrary to a similar results by [Koyejo et al \(2014\)](#), Theorem 3 does not require continuity of the cumulative distribution of  $\eta(x)$  around  $\theta^*$ .

## 4 Proof of Lemma 1

The proof can be skipped without affecting the flow of later sections. The proof consists of two steps. First, we bound the  $\Psi$ -regret of any classifier  $h$  by its cost-sensitive classification regret (introduced below). Next, we show that there exists a threshold  $\theta^*$ , such that for any  $f$ , the cost-sensitive classification regret of  $h_{f,\theta^*}$  is upperbounded by the  $\ell$ -regret of  $f$ . These two steps will be formalized as Proposition 4 and Proposition 5.

Given a real number  $\alpha \in [0, 1]$ , define a *cost-sensitive classification loss*  $\ell_\alpha: \{-1, 1\} \times \{-1, 1\} \rightarrow \mathbb{R}_+$  as:

$$\ell_\alpha(y, \hat{y}) = \alpha \mathbb{I}[y = -1] \mathbb{I}[\hat{y} = 1] + (1 - \alpha) \mathbb{I}[y = 1] \mathbb{I}[\hat{y} = -1].$$

The cost-sensitive loss assigns different costs of misclassification for positive and negative labels. Given classifier  $h$ , the *cost-sensitive risk* of  $h$  is:

$$\begin{aligned} \text{Risk}_\alpha(h) &= \mathbb{E}_{(x,y)}[\ell_\alpha(y, h(x))] \\ &= \alpha \text{FP}(h) + (1 - \alpha) \text{FN}(h), \end{aligned}$$

and the *cost-sensitive regret* is:

$$\text{Reg}_\alpha(h) = \text{Risk}_\alpha(h) - \text{Risk}_\alpha(h_\alpha^*),$$

where  $h_\alpha^* = \arg\min_h \text{Risk}_\alpha(h)$ . We now show the following two results:

**Proposition 4.** *Let  $\Psi$  satisfy the assumptions from Lemma 1. Define:*

$$\alpha = \frac{\Psi^* b_1 - a_1}{\Psi^*(b_1 + b_2) - (a_1 + a_2)}. \quad (5)$$

*Then,  $\alpha \in [0, 1]$  and for any classifier  $h$ ,*

$$\text{Reg}_\Psi(h) \leq C \text{Reg}_\alpha(h),$$

*where  $C$  is defined as in the content of Lemma 1.*

*Proof.* The proof generalizes the proof of Proposition 6 from Parambath et al (2014), which concerned the special case of  $F_\beta$ -measure. For the sake of clarity, we use a shorthand notation  $\Psi = \Psi(h)$ ,  $\Psi^* = \Psi(h_\Psi^*)$ ,  $\text{FP} = \text{FP}(h)$ ,  $\text{FN} = \text{FN}(h)$ ,  $A = a_0 + a_1 \text{FP} + a_2 \text{FN}$ ,  $B = b_0 + b_1 \text{FP} + b_2 \text{FN}$  for the numerator and denominator of  $\Psi(h)$ , and analogously  $\text{FP}^*$ ,  $\text{FN}^*$ ,  $A^*$  and  $B^*$  for  $\Psi(h_\Psi^*)$ . In this notation:

$$\begin{aligned} \text{Reg}_\Psi(h) &= \Psi^* - \Psi = \frac{\Psi^* B - A}{B} \\ &= \frac{\Psi^* B - A - \overbrace{(\Psi^* B^* - A^*)}^{=0}}{B} \\ &= \frac{\Psi^*(B - B^*) - (A - A^*)}{B} \\ &= \frac{(\Psi^* b_1 - a_1)(\text{FP} - \text{FP}^*) + (\Psi^* b_2 - a_2)(\text{FN} - \text{FN}^*)}{B} \\ &\leq \frac{(\Psi^* b_1 - a_1)(\text{FP} - \text{FP}^*) + (\Psi^* b_2 - a_2)(\text{FN} - \text{FN}^*)}{\gamma}, \end{aligned} \quad (6)$$

where the last inequality follows from  $B \geq \gamma$  (assumption) and the fact that  $\text{Reg}_\Psi(h) \geq 0$  for any  $h$ . Since  $\Psi$  is non-increasing in  $\text{FP}$  and  $\text{FN}$ , we have

$$\frac{\partial \Psi^*}{\partial \text{FP}^*} = \frac{a_1 B^* - b_1 A^*}{(B^*)^2} = \frac{a_1 - b_1 \Psi^*}{B^*} \leq 0,$$

and similarly  $\frac{\partial \Psi^*}{\partial \text{FN}^*} = \frac{a_2 - b_2 \Psi^*}{B^*} \leq 0$ . This and the assumption  $B^* \geq \gamma$  implies that both  $\Psi^* b_1 - a_1$  and  $\Psi^* b_2 - a_2$  are non-negative, so can be interpreted as misclassification costs. If we normalize the costs by defining:

$$\alpha = \frac{\Psi^* b_1 - a_1}{\Psi^* (b_1 + b_2) - (a_1 + a_2)},$$

then (6) implies:

$$\begin{aligned} \text{Reg}_\Psi(h) &\leq C (\text{Risk}_\alpha(h) - \text{Risk}_\alpha(h_\Psi^*)) \\ &\leq C (\text{Risk}_\alpha(h) - \text{Risk}_\alpha(h_\alpha^*)) = C \text{Reg}_\alpha(h). \end{aligned}$$

□

**Proposition 5.** *For any real-valued function  $f: X \rightarrow \mathbb{R}$  any  $\lambda$ -strongly proper composite loss  $\ell$  with link function  $\psi$ , and any  $\alpha \in [0, 1]$ :*

$$\text{Reg}_\alpha(h_{f, \theta^*}) \leq \sqrt{\frac{2}{\lambda}} \sqrt{\text{Reg}_\ell(f)}, \quad (7)$$

where  $\theta^* = \psi(\alpha)$ .

*Proof.* First, we will show that (7) holds *conditionally* for every  $x$ . To this end, we fix  $x$  and deal with  $h(x) \in \{-1, 1\}$ ,  $f(x) \in \mathbb{R}$  and  $\eta(x) \in [0, 1]$ , using a shorthand notation  $h, f, \eta$ .

Given  $\eta \in [0, 1]$  and  $h \in \{-1, 1\}$ , define the *conditional cost-sensitive risk* as:

$$\text{risk}_\alpha(\eta, h) = \alpha(1 - \eta)\mathbb{I}[h = 1] + (1 - \alpha)\eta\mathbb{I}[h = -1].$$

Let  $h_\alpha^* = \text{argmin}_h \text{risk}_\alpha(\eta, h)$ . It can be easily verified that:

$$h_\alpha^* = \text{sgn}(\eta - \alpha). \quad (8)$$

Define the *conditional cost-sensitive regret* as

$$\text{reg}_\alpha(\eta, h) = \text{risk}_\alpha(\eta, h) - \text{risk}_\alpha(\eta, h_\alpha^*).$$

Note that if  $h = h_\alpha^*$ , then  $\text{reg}_\alpha(\eta, h) = 0$ . Otherwise,  $\text{reg}_\alpha(\eta, h) = |\eta - \alpha|$ , so that:

$$\text{reg}_\alpha(\eta, h) = \mathbb{I}[h \neq h_\alpha^*] |\eta - \alpha|.$$

Now assume  $h = \text{sgn}(\hat{\eta} - \alpha)$  for some  $\hat{\eta}$ , i.e.,  $h$  is of the same form as  $h_\alpha^*$  in (8), with  $\eta$  replaced by  $\hat{\eta}$ . We show that for such  $h$ ,

$$\text{reg}_\alpha(\eta, h) \leq |\eta - \hat{\eta}|. \quad (9)$$

This statement trivially holds when  $h = h_\alpha^*$ . If  $h \neq h_\alpha^*$ , then  $\eta$  and  $\hat{\eta}$  are on the opposite sides of  $\alpha$  (i.e. either  $\eta \geq \alpha$  and  $\hat{\eta} < \alpha$  or  $\eta < \alpha$  and  $\hat{\eta} \geq \alpha$ ), hence  $|\eta - \alpha| \leq |\eta - \hat{\eta}|$ , which proves (9).

Now, we set the threshold to  $\theta^* = \psi(\alpha)$ , so that given  $f \in \mathbb{R}$ ,

$$h_{f, \theta^*} = \text{sgn}(f - \theta^*) = \text{sgn}(f - \psi(\alpha)) = \text{sgn}(\psi^{-1}(f) - \alpha),$$

due to strict monotonicity of  $\psi$ . Using (9) with  $h = h_{f, \theta^*}$  and  $\hat{\eta} = \psi^{-1}(f)$  gives:

$$\begin{aligned} \text{reg}_\alpha(\eta, h_{f, \theta^*}) &\leq |\eta - \psi^{-1}(f)| = \sqrt{(\eta - \psi^{-1}(f))^2} \\ &\leq \sqrt{\frac{2}{\lambda}} \sqrt{\text{reg}_\ell(\eta, f)}, \end{aligned} \quad (10)$$

and the last inequality follows from strong properness (2).

To prove the unconditional statement (7), we take expectation with respect to  $x$  on both sides of (10):

$$\begin{aligned}
\text{Reg}_\alpha(h_{f,\theta^*}) &= \mathbb{E}_x [\text{reg}_\alpha(\eta, h_{f,\theta^*}(x))] \\
\text{(by (10))} \quad &\leq \sqrt{\frac{2}{\lambda}} \mathbb{E}_x [\sqrt{\text{reg}_\ell(\eta(x), f(x))}] \\
&\leq \sqrt{\frac{2}{\lambda}} \sqrt{\mathbb{E}_x [\text{reg}_\ell(\eta(x), f(x))]} \\
&= \sqrt{\frac{2}{\lambda}} \sqrt{\text{Reg}_\ell(f)}, \tag{11}
\end{aligned}$$

where the second inequality is from Jensen's inequality applied to the concave function  $x \mapsto \sqrt{x}$ .

We note that derivation of (9) follows the steps of the proof of Lemma 4 in Menon et al (2013), while (10) and (11) were shown in the proof of Theorem 13 by Agarwal (2014). Hence, the proof is essentially a combination of existing results, which are rederived here for the sake of completeness.  $\square$

*Proof of Lemma 1.* Lemma 1 immediately follows from Proposition 4 and Proposition 5.  $\square$

Note that the proof actually specifies the exact value of the universal threshold,  $\theta^* = \psi(\alpha)$ , where  $\alpha$  is given by (5).

The bound in Lemma 1 is unimprovable in a sense that there exist  $f$ ,  $\Psi$ ,  $\ell$ , and distribution  $\Pr(x, y)$  for which the bound is tight. To see this, take, for instance, squared error loss  $\ell(y, f) = (y - f)^2$  and classification accuracy metric  $\Psi(\text{FP}, \text{FN}) = 1 - \text{FP} - \text{FN}$ . The constants in Lemma 1 are equal to  $\gamma = 1$ ,  $C = 2$ , and  $\lambda = 8$  (see Table 1), while the optimal threshold is  $\theta^* = 0$ . The bound then simplifies to

$$\text{Reg}_{0/1}(\text{sgn}(f)) \leq \sqrt{\text{Reg}_{\text{sqr}}(f)},$$

which is known to be tight (Bartlett et al, 2006).

## 5 Multilabel classification

In multilabel classification (Dembczyński et al, 2012; Parambath et al, 2014; Koyejo et al, 2015), the goal is, given an input (feature vector)  $x \in X$ , to simultaneously predict the subset  $L \subseteq \mathcal{L}$  of the set of  $m$  labels  $\mathcal{L} = \{\sigma_1, \dots, \sigma_m\}$ . The subset  $L$  is often called the set of relevant (positive) labels, while the complement  $\mathcal{L} \setminus L$  is considered as irrelevant (negative) for  $x$ . We identify a set  $L$  of relevant labels with a vector  $\mathbf{y} = (y_1, y_2, \dots, y_m)$ ,  $y_i \in \{-1, 1\}$ , in which  $y_i = 1$  iff  $\sigma_i \in L$ . We assume observations  $(x, \mathbf{y})$  are generated i.i.d. according to  $\Pr(x, \mathbf{y})$  (note that the labels are not assumed to be independent). A *multilabel classifier*:

$$\mathbf{h}(x) = (h_1(x), h_2(x), \dots, h_m(x)),$$

is a mapping  $\mathbf{h}: X \rightarrow \{-1, 1\}^m$ , which assigns a (predicted) label subset to each instance  $x \in X$ . For any  $i = 1, \dots, m$ , the function  $h_i(x)$  is thus a binary classifier, which can be evaluated by means of  $\text{TP}_i(h_i)$ ,  $\text{FP}_i(h_i)$ ,  $\text{TN}_i(h_i)$  and  $\text{FN}_i(h_i)$ , which are true/false positives/negatives defined with respect to label  $y_i$ , e.g.  $\text{FP}_i(h_i) = \Pr(h_i(x) = 1 \wedge y_i = -1)$ .

Let  $f_1, \dots, f_m$  be a set of real-valued functions  $f_i: X \rightarrow \mathbb{R}$ ,  $i = 1, \dots, m$ , and let  $\ell$  be a  $\lambda$ -strongly proper composite loss for binary classification. For each  $i = 1, \dots, m$ , we let  $\text{Risk}_\ell^i(f_i)$  and  $\text{Reg}_\ell^i(f_i)$  denote the  $\ell$ -risk and the  $\ell$ -regret of function  $f_i$  with respect to label  $y_i$ :

$$\text{Risk}_\ell^i(f_i) = \mathbb{E}_{(x, y_i)} [\ell(y_i, f_i(x))], \quad \text{Reg}_\ell^i(f_i) = \text{Risk}_\ell^i(f_i) - \min_f \text{Risk}_\ell^i(f).$$

Note that the problem has been decomposed into  $m$  independent binary problems and the functions can be obtained by training  $m$  independent real-valued binary classifiers by minimizing loss  $\ell$  on the training sample, one for each out of  $m$  labels.

What follows next depends on the way in which the binary classification performance metric is applied in the multilabel setting. We consider two ways of turning binary classification metric into multilabel metric: the macro-averaging and the micro-averaging (Manning et al, 2008; Parambath et al, 2014; Koyejo et al, 2015).

## 5.1 Macro-averaging

Given a binary classification performance metric  $\Psi(h) = \Psi(\text{FP}(h), \text{FN}(h))$ , and a multilabel classifier  $\mathbf{h}$ , we define the macro-averaged metric  $\Psi_{\text{macro}}(\mathbf{h})$  as:

$$\Psi_{\text{macro}}(\mathbf{h}) = \frac{1}{m} \sum_{i=1}^m \Psi(h_i) = \frac{1}{m} \sum_{i=1}^m \Psi(\text{FP}_i(h_i), \text{FN}_i(h_i)).$$

The macro-averaging is thus based on first computing the performance metric separately for each label, and then averaging the metrics over the labels. The  $\Psi_{\text{macro}}$ -regret is then defined as:

$$\text{Reg}_{\Psi_{\text{macro}}}(\mathbf{h}) = \Psi_{\text{macro}}(\mathbf{h}_{\Psi}^*) - \Psi_{\text{macro}}(\mathbf{h}) = \frac{1}{m} \sum_{i=1}^m \left( \Psi(h_{\Psi,i}^*) - \Psi(h_i) \right),$$

where  $\mathbf{h}_{\Psi}^* = (h_{\Psi,1}^*, \dots, h_{\Psi,m}^*)$  is the  $\Psi$ -optimal multilabel classifier:

$$h_{\Psi,i}^* = \underset{h}{\operatorname{argmax}} \Psi(\text{FP}_i(h), \text{FN}_i(h)), \quad i = 1, \dots, m.$$

Since the regret decomposes into a weighted sum, it is straightforward to apply previously derived bound to obtain a regret bound for macro-averaged performance metric.

**Theorem 6.** *Let  $\Psi(\text{FP}, \text{FN})$  and  $\ell$  satisfy the assumptions of Lemma 1. For a set of  $m$  real-valued functions  $\{f_i: X \rightarrow \mathbb{R}\}_{i=1}^m$ , let  $\theta_{f_i}^* = \underset{\theta}{\operatorname{argmax}} \Psi(h_{f_i, \theta})$  for each  $i = 1, \dots, m$ . Then the classifier  $\mathbf{h}$  defined as:*

$$\mathbf{h} = (h_{f_1, \theta_{f_1}^*}, h_{f_2, \theta_{f_2}^*}, \dots, h_{f_m, \theta_{f_m}^*}),$$

*achieves the following bound on its  $\Psi_{\text{macro}}$ -regret:*

$$\text{Reg}_{\Psi_{\text{macro}}}(\mathbf{h}) \leq \sqrt{\frac{2}{\lambda}} \frac{1}{m} \sum_{i=1}^m C_i \sqrt{\text{Reg}_{\ell}^i(f_i)},$$

where  $C_i = \frac{1}{\gamma} \left( \Psi(h_{\Psi,i}^*)(b_1 + b_2) - (a_1 + a_2) \right)$ ,  $i = 1, \dots, m$ .

*Proof.* The theorem follows from applying Theorem 2 once for each label, and then averaging the bounds over the labels.  $\square$

Theorem 6 suggests a straightforward decomposition into  $m$  independent binary classification problems, one for each label  $y_1, \dots, y_m$ , and running (independently for each problem) the two-step procedure described in Section 3: For  $i = 1, \dots, m$ , we learn a function  $f_i$  with small  $\ell$ -regret with respect to label  $y_i$ , and tune the threshold  $\theta_{f_i}^*$  to optimize  $\Psi(h_{f_i, \theta})$  (similarly as in the binary classification case, one can show that tuning the threshold on a separate validation sample is sufficient). Due to decomposition of  $\Psi_{\text{macro}}$  into the sum over the labels, this simple procedure turns out to be sufficient. As we shall see, the case of micro-averaging becomes more interesting.

## 5.2 Micro-averaging

Given a binary classification performance metrics  $\Psi(h) = \Psi(\text{FP}(h), \text{FN}(h))$ , and a multilabel classifier  $\mathbf{h}$ , we define the micro-averaged metric  $\Psi_{\text{micro}}(\mathbf{h})$  as:

$$\Psi_{\text{micro}}(\mathbf{h}) = \Psi(\overline{\text{FP}}(\mathbf{h}), \overline{\text{FN}}(\mathbf{h})),$$

where:

$$\overline{\text{FP}}(\mathbf{h}) = \frac{1}{m} \sum_{i=1}^m \text{FP}_i(h_i), \quad \overline{\text{FN}}(\mathbf{h}) = \frac{1}{m} \sum_{i=1}^m \text{FN}_i(h_i).$$

Thus, in the micro-averaging, the false positives and false negatives are first averaged over the labels, and then the performance metric is calculated on these averaged quantities. The  $\Psi_{\text{micro}}$ -regret:

$$\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}) = \Psi_{\text{micro}}(\mathbf{h}_{\Psi}^*) - \Psi_{\text{micro}}(\mathbf{h}), \quad \text{where } \mathbf{h}_{\Psi}^* = \underset{\mathbf{h}}{\text{argmax}} \Psi_{\text{micro}}(\mathbf{h}),$$

does not decompose into the sum over labels anymore. However, we are still able to obtain a regret bound, reusing the techniques from Section 4, and, interestingly, this time only a single threshold needs to be tuned and is shared among all labels.<sup>3</sup>

**Theorem 7.** *Let  $\Psi(\text{FP}, \text{FN})$  and  $\ell$  satisfy the assumptions of Lemma 1. For a set of  $m$  real-valued functions  $\{f_i: X \rightarrow \mathbb{R}\}_{i=1}^m$ , let  $\theta_f^* = \text{argmax}_{\theta} \Psi_{\text{micro}}(\mathbf{h}_{f,\theta})$ , where:*

$$\mathbf{h}_{f,\theta} = (h_{f_1,\theta}, h_{f_2,\theta}, \dots, h_{f_m,\theta}).$$

*Then, the classifier  $\mathbf{h}_{f,\theta_f^*} = (h_{f_1,\theta_f^*}, \dots, h_{f_m,\theta_f^*})$  achieves the following bound on its  $\Psi_{\text{micro}}$ -regret:*

$$\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}_{f,\theta_f^*}) \leq \sqrt{\frac{2}{\lambda}} \frac{C}{m} \sum_{i=1}^m \sqrt{\text{Reg}_{\ell}^i(f_i)},$$

where  $C = \frac{1}{\gamma} (\Psi_{\text{micro}}(\mathbf{h}_{\Psi}^*)(b_1 + b_2) - (a_1 + a_2))$ .

*Proof.* The proof follows closely the proof of Lemma 1. In fact, only Proposition 4 requires modifications, which are given below. Take any real values  $\text{FP}, \text{FN}$  and  $\text{FP}^*, \text{FN}^*$  (to be specified later) in the domain of  $\Psi$ , such that:

$$\Psi(\text{FP}^*, \text{FN}^*) - \Psi(\text{FP}, \text{FN}) \geq 0. \tag{12}$$

Using exactly the same steps as in the derivation (6), we obtain:

$$\Psi(\text{FP}^*, \text{FN}^*) - \Psi(\text{FP}, \text{FN}) \leq C (\alpha(\text{FP} - \text{FP}^*) + (1 - \alpha)(\text{FN} - \text{FN}^*)),$$

where:

$$C = \frac{1}{\gamma} (\Psi(\text{FP}^*, \text{FN}^*)(b_1 + b_2) - (a_1 + a_2)),$$

$$\alpha = \frac{\Psi(\text{FP}^*, \text{FN}^*)b_1 - a_1}{\Psi(\text{FP}^*, \text{FN}^*)(b_1 + b_2) - (a_1 + a_2)}.$$

Now, we take:  $\text{FP}^* = \overline{\text{FP}}(\mathbf{h}_{\Psi}^*)$ ,  $\text{FN}^* = \overline{\text{FN}}(\mathbf{h}_{\Psi}^*)$ ,  $\text{FP} = \overline{\text{FP}}(\mathbf{h})$  and  $\text{FN} = \overline{\text{FN}}(\mathbf{h})$  for some  $\mathbf{h}$ . Hence, (12) is clearly satisfied as its left-hand side is just the  $\Psi_{\text{micro}}$ -regret,  $\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h})$ . This means that for any

---

<sup>3</sup>The fact that a single threshold is sufficient for consistency of micro-averaged performance measures was already noticed by Koyejo et al (2015).

multilabel classifier  $\mathbf{h}$ :

$$\begin{aligned}
\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}) &\leq C \left( \alpha(\overline{\text{FP}}(\mathbf{h}) - \overline{\text{FP}}(\mathbf{h}_{\Psi}^*)) + (1 - \alpha)(\overline{\text{FN}}(\mathbf{h}) - \overline{\text{FN}}(\mathbf{h}_{\Psi}^*)) \right) \\
&= \frac{C}{m} \sum_{i=1}^m \alpha(\text{FP}_i(h_i) - \text{FP}_i(h_{\Psi,i}^*)) + (1 - \alpha)(\text{FN}_i(h_i) - \text{FN}_i(h_{\Psi,i}^*)) \\
&= \frac{C}{m} \sum_{i=1}^m \left( \text{Risk}_{\alpha}^i(h_i) - \text{Risk}_{\alpha}^i(h_{\Psi,i}^*) \right) \\
&\leq \frac{C}{m} \sum_{i=1}^m \text{Reg}_{\alpha}^i(h_i),
\end{aligned}$$

where  $\text{Risk}_{\alpha}^i(h_i)$  and  $\text{Reg}_{\alpha}^i(h_i)$  are the cost-sensitive risk and the cost sensitive regret defined with respect to label  $y_i$ :

$$\text{Risk}_{\alpha}^i(h_i) = \mathbb{E}_{(x,y_i)}[\ell_{\alpha}(y_i, h_i(x))], \quad \text{Reg}_{\alpha}^i(h_i) = \text{Risk}_{\alpha}^i(h_i) - \min_h \text{Risk}_{\alpha}^i(h).$$

If we now take  $h_i = h_{f,\theta^*}$ , where  $\theta^* = \psi(\alpha)$ ,  $\psi$  being the link function of the loss, Proposition 5 (applied for each  $i = 1, \dots, m$  separately) implies:

$$\text{Reg}_{\alpha}^i(h_{f,\theta^*}) \leq \sqrt{\frac{2}{\lambda}} \sqrt{\text{Reg}_{\ell}^i(f_i)}.$$

Together, this gives:

$$\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}_{f,\theta^*}) \leq \sqrt{\frac{2}{\lambda}} \frac{C}{m} \sum_{i=1}^m \sqrt{\text{Reg}_{\ell}^i(f_i)}.$$

The theorem now follows by noticing that:

$$\theta_f^* = \underset{\theta}{\text{argmax}} \Psi_{\text{micro}}(\mathbf{h}_{f,\theta}) = \underset{\theta}{\text{argmin}} \text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}_{f,\theta}),$$

and thus  $\text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}_{f,\theta_f^*}) \leq \text{Reg}_{\Psi_{\text{micro}}}(\mathbf{h}_{f,\theta^*})$ .  $\square$

Theorem 7 suggests a decomposition into  $m$  independent binary classification problems, one for each label  $y_1, \dots, y_m$ , and training  $m$  real-valued classifiers  $f_1, \dots, f_m$  with small  $\ell$ -regret on the corresponding label. Then, however, contrary to macro-averaging, a single threshold, shared among all labels, is tuned by optimizing  $\Psi_{\text{micro}}$  on a separate validation sample.

## 6 Empirical results

We perform experiments on synthetic and benchmark data to empirically study the two-step procedure analyzed in the previous sections. To this end, we minimize a surrogate loss in the first step to obtain a real-valued function  $f$ , and in the second step, we tune a threshold  $\hat{\theta}$  on a separate validation set to optimize a given performance metric. We use logistic loss in this procedure as a surrogate loss. Recall that logistic loss is 4-strongly proper composite (see Table 2). We compare its performance with *hinge loss*, which is even *not* a proper composite function. As our task performance metrics, we take the F-measure ( $F_{\beta}$ -measure with  $\beta = 1$ ) and the AM measure (which is a special case of Weighted Accuracy with weights  $w_1 = P$  and  $w_2 = 1 - P$ ). We could also use the Jaccard similarity coefficient; it turns out, however, that the threshold optimized for the F-measure coincides with the optimal threshold for the Jaccard similarity coefficient (this is because the Jaccard similarity coefficient is strictly monotonic in the F-measure and vice versa), so the latter measure does not give anything substantially different than the F-measure.

The experiments on benchmark data are split into two parts. The first part concerns binary classification problems, while the second part multi-label classification.



The purpose of this study is *not* about comparing the two-step approach with alternative methods; this has already been done in the previous work on the subject, see, e.g., (Nan et al, 2012; Parambath et al, 2014). We also note that similar experiments have been performed in the cited papers on the statistical consistency of generalized performance metrics (Koyejo et al, 2014; Narasimhan et al, 2014; Parambath et al, 2014; Koyejo et al, 2015). Therefore, we unavoidably repeat some of the results obtained therein, but the main novelty of the experiments reported here is that we emphasize the difference between strongly proper composite losses and non-proper losses.

## 6.1 Synthetic data

We performed two experiments on synthetic data. The first experiment deals with a discrete domain in which we learn within a class of all possible classifiers. The second experiment concerns continuous domain in which we learn within a restricted class of linear functions.

**First experiment.** We let the input domain  $X$  to be a finite set, consisting of 25 elements,  $X = \{1, 2, \dots, 25\}$ , and take  $\Pr(x)$  to be uniform over  $X$ , i.e.  $\Pr(x = i) = 1/25$ . For each  $x \in X$ , we randomly draw a value of  $\eta(x)$  from the uniform distribution on the interval  $[0, 1]$ . In the first step, we take an algorithm which minimizes a given surrogate loss  $\ell$  within the class of *all* function  $f: X \rightarrow \mathbb{R}$ . Hence, given the training data of size  $n$ , the algorithm computes the empirical minimizer of surrogate loss  $\ell$  independently for each  $x$ . As surrogate losses, we use logistic and hinge loss. In the second step, we tune the threshold  $\theta$  on a separate validation set, also of size  $n$ . For each  $n$ , we repeat the procedure 100,000 times, averaging over samples and over models (different random choices of  $\eta(x)$ ). We start with  $n = 100$  and increase the number of training examples up to  $n = 10,000$ . The  $\ell$ -regret and  $\Psi$ -regret can be easily computed, as the distribution is known and  $X$  is discrete.

The results are given in Fig. 1. The  $\ell$ -regret goes down to zero for both surrogate losses, which is expected, since this is the objective function minimized by the algorithm. Minimization of logistic loss (left plot) gives vanishing  $\Psi$ -regret for both the F-measure and the AM measure, as predicted by Theorem 2. In contrast, minimization of the hinge loss (right plot) is suboptimal for both task metrics and gives non-zero  $\Psi$ -regret even in the limit  $n \rightarrow \infty$ . This behavior can easily be explained by the fact that hinge loss is not a proper (composite) loss: the risk minimizer for hinge loss is given by  $f_\ell^*(x) = \text{sgn}(\eta(x) - 1/2)$  (Bartlett et al, 2006). Hence, the hinge loss minimizer is already a threshold function on  $\eta(x)$ , with the threshold value set to  $1/2$ . If, for a given performance metric  $\Psi$ , the optimal threshold  $\theta^*$  is different than  $1/2$ , the hinge loss minimizer will necessarily have suboptimal  $\Psi$ -risk. This is clearly visible for the F-measure. The better result on the AM measure is explained by the fact that the average optimal threshold over all models is  $0.5$  for this measure, so the minimizer of hinge loss is not that far from the minimizer of AM measure.

**Second experiment.** We take  $X = \mathbb{R}^2$  and generate  $x \in X$  from a standard Gaussian distribution. We use a logistic model of the form  $\eta(x) = \frac{1}{1 + \exp(-a_0 - a^\top x)}$ . The weights  $a = (a_1, a_2)$  and  $a_0$  are also drawn from a standard Gaussian. For a given model (set of weights), we take training sets of increasing size from  $n = 100$  up to  $n = 3000$ , using 20 different sets for each  $n$ . We also generate one test set of size 100,000. For each  $n$ , we use  $2/3$  of the training data to learn a linear model  $f(x) = w_0 + w^\top x$ , using either support vector machines (SVM, with linear kernel) or logistic regression (LR). We use implementation of these algorithms from the LibLinear package (Fan et al, 2008).<sup>4</sup> The remaining  $1/3$  of the training data is used for tuning the threshold. We average the results over 20 different models.

The results are given in Fig. 2. As before, we plot the average  $\ell$ -regret for logistic and hinge loss, and  $\Psi$ -regret for the F-measure and the AM measure. The results obtained for LR (logistic loss minimizer) agree with our theoretical analysis: the  $\ell$ -regret and  $\Psi$ -regret with respect to both F-measure and AM measure go to zero. This is expected, as the data generating model is a linear logistic model (so that the risk minimizer for logistic loss is a linear function), and thus coincides with a class of functions over which we optimize. The situation is different for SVM (hinge loss minimizer). Firstly, the  $\ell$ -regret for hinge loss does not converge to

<sup>4</sup>Software available at <http://www.csie.ntu.edu.tw/~cjlin/liblinear>

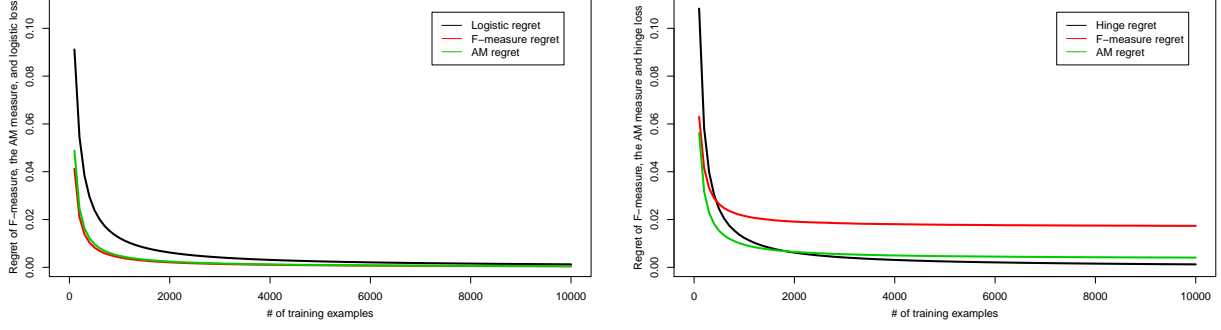


Figure 1: Regret (averaged over 100,000 repetitions) on the discrete synthetic model as a function of the number of training examples. Left panel: logistic loss is used as a surrogate loss. Right panel: hinge loss is used as surrogate loss.

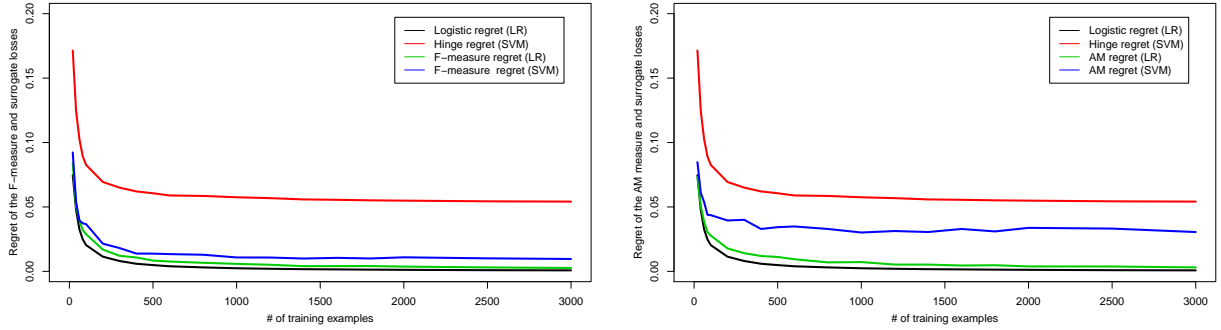


Figure 2: Regret (averaged over  $20 \times 20 = 400$  repetitions) on the logistic model as a function of the number of training examples. Left panel: regret with respect to the F-measure and surrogate losses. Right panel: regret with respect to the AM measure and surrogate losses.

zero. This is because the risk minimizer for hinge loss is a threshold function  $\text{sgn}(\eta(x) - 1/2)$ , and it is not possible to approximate such a function with linear model  $f(x) = w_0 + w^\top x$ . Hence, even when  $n \rightarrow \infty$ , the empirical hinge loss minimizer (SVM) does not converge to the risk minimizer. This behavior, however, can be *advantageous* for SVM in terms of the task performance measures. This is because the risk minimizer for hinge loss, a threshold function on  $\eta(x)$  with the threshold value  $1/2$ , will perform poorly, for example, in terms of the F-measure and AM measure, for which the optimal threshold  $\theta^*$  is usually very different from  $1/2$ . In turn, the linear model constraint will prevent convergence to the risk minimizer, and the resulting linear function  $f(x) = w_0 + w^\top x$  will often be close to some reversible function of  $\eta(x)$ ; hence after tuning the threshold, we will often end up close to the minimizer of a given task performance measure. This is seen for the F-measure on the left panel in Fig. 2. In this case, the F-regret of SVM gets quite close to zero, but is still worse than LR. The non-vanishing regret is mainly caused by the fact that for some models with imbalanced class priors, SVM reduce weights  $w$  to zero and sets the intercept  $w_0$  to 1 or  $-1$ , predicting the same value for all  $x \in X$  (this is not caused by a software problem, it is how the empirical loss minimizer behaves). Interestingly, the F-measure is only slightly affected by this pathological behavior of empirical hinge loss minimizer. In turn, the AM measure, for which the plots are drawn in the right panel in Fig. 2, is not robust against this behavior of SVM: predicting the majority class actually results in the value of AM measure equal to  $1/2$ , a very poor performance, which is on the same level as random classifier.

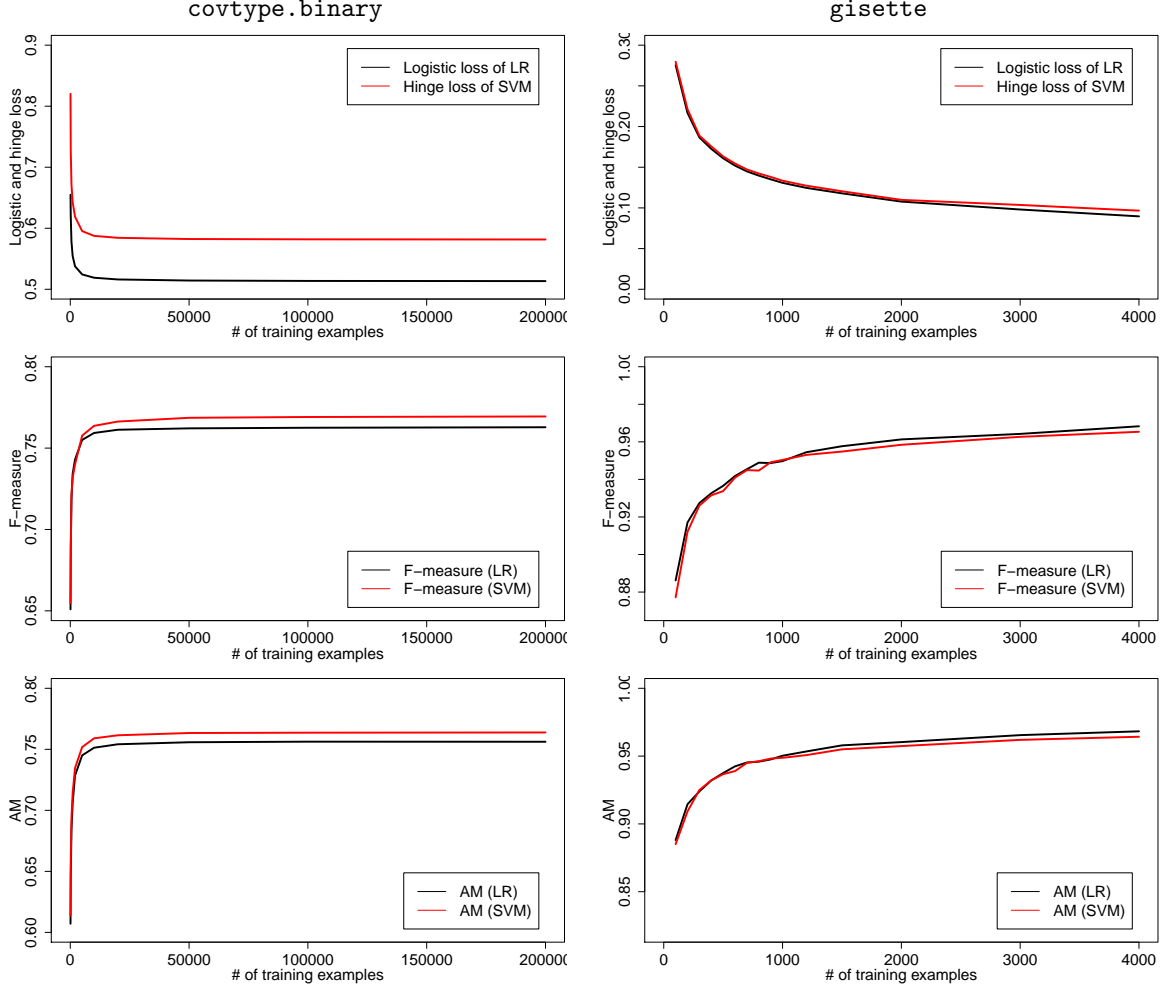


Figure 3: Average test set performance on benchmark data sets as a function of the number of training examples. Left panel: **covtype** dataset. Right panel: the **gisette** dataset. The top plots show logistic and hinge loss, the center plots show the F-measure, the bottom plots show the AM measure.

## 6.2 Benchmark data for binary classification

The next experiment is performed on two binary benchmark datasets,<sup>5</sup> described in Table 4. We randomly take out a test set of size 181,012 for **covtype**, and of size 3,000 for **gisette**. We use the remaining examples for training. As before, we incrementally increase the size of the training set. We use 2/3 of training examples for learning linear model with SVM or LR, and the rest for tuning the threshold. We repeat the experiment (random train/validation/test split) 20 times. The results are plotted in Fig 3. Since the data distribution is unknown, we are unable to compute the risk minimizers, hence we plot the average loss/metric on the test set rather than the regret. The results show that SVM perform better on the **covtype** dataset, while LR performs better on the **gisette** dataset. However, there is very little difference in performance of SVM and LR in terms of the F-measure and the AM measure on these data sets. We suspect this is due to the fact that  $\eta(x)$  function is very different from linear for these problems, so that neither LR nor SVM converge to the  $\ell$ -risk minimizer, and Theorem 2 does not apply. Further studies would be required to understand the

<sup>5</sup> Datasets are taken from LibSVM repository: <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>

dataset	#examples	#features
covtype	581,012	54
gisette	7,000	5,000

Table 4: Basic statistics for binary classification benchmark datasets

data set	# labels	# training examples	# test examples	#features
scene	6	1211	1169	294
yeast	14	1500	917	103
mediamill	101	30993	12914	120

Table 5: Basic statistics for multi-label benchmark data sets

behavior of surrogate losses in this case.

### 6.3 Benchmark data for multi-label classification

In the last experiment we use three multi-label benchmark data sets.<sup>6</sup> Table 5 provides a summary of basic statistics of these datasets. The aim of the experiment is to verify the theoretical results in Section 5 on learning the micro- and macro-averaged performance metrics. We use the F-measure and the AM-measure as in previous experiments.

The data sets are already split into the training and testing parts. As before we train a linear model using either SVM or LR on 2/3 of training examples. The rest of training data is used for tuning the threshold. For optimizing macro-averaged measures, we tune the threshold separately for each label. This approach agrees with our analysis given in Section 5.1. For micro-averaging, we tune a common threshold for all labels: we simply collect predictions for all labels and find the best threshold using these values. This approach is justified by the theoretical analysis in Section 5.2. Hence, the only difference between micro- and macro-versions of the algorithms is whether a single or multiple thresholds are tuned. In total we use 8 algorithms: two learning algorithms (LR/SVM), two performance measures (F/AM), and two types of averaging (Macro/Micro). Note that our experiments include evaluating algorithms tuned for macro-averaging in terms of micro-averaged metrics, and vice versa. The goal of such cross-analysis is to determine the impact of threshold sharing for both averaging schemes. As before, we incrementally increase the size of the training set and repeat training and threshold tuning 20 times (we use random draws of training instances into the proper training and the validation parts; the test set is always the same, as originally specified for each data set). The results are given in Fig 4.

The plots generally agree with the conclusions coming from the theoretical analysis, with some intriguing exceptions, however. As expected, LR tuned for a given performance metric gets the best result with respect to that metric in most of the cases. For the **scene** data set, however, the methods tuned for the micro-averaged metrics (single threshold shared among labels) outperform the ones tuned for macro-averaged metrics (separate thresholds tuned for each label), even when evaluated in terms of macro-averaged metrics. A similar result has been obtained by Koyejo et al (2015). It seems that tuning a single threshold shared among all labels can lead to a more stable solution that is less prone to overfitting, even though it is not the optimal thing to do for macro-averaged measures. We further report that, interestingly, SVM outperform LR in terms of Macro-F on **mediamill** and this is the only case in which SVM get a better result than LR.

<sup>6</sup>Datasets are taken from LibSVM repository: <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>

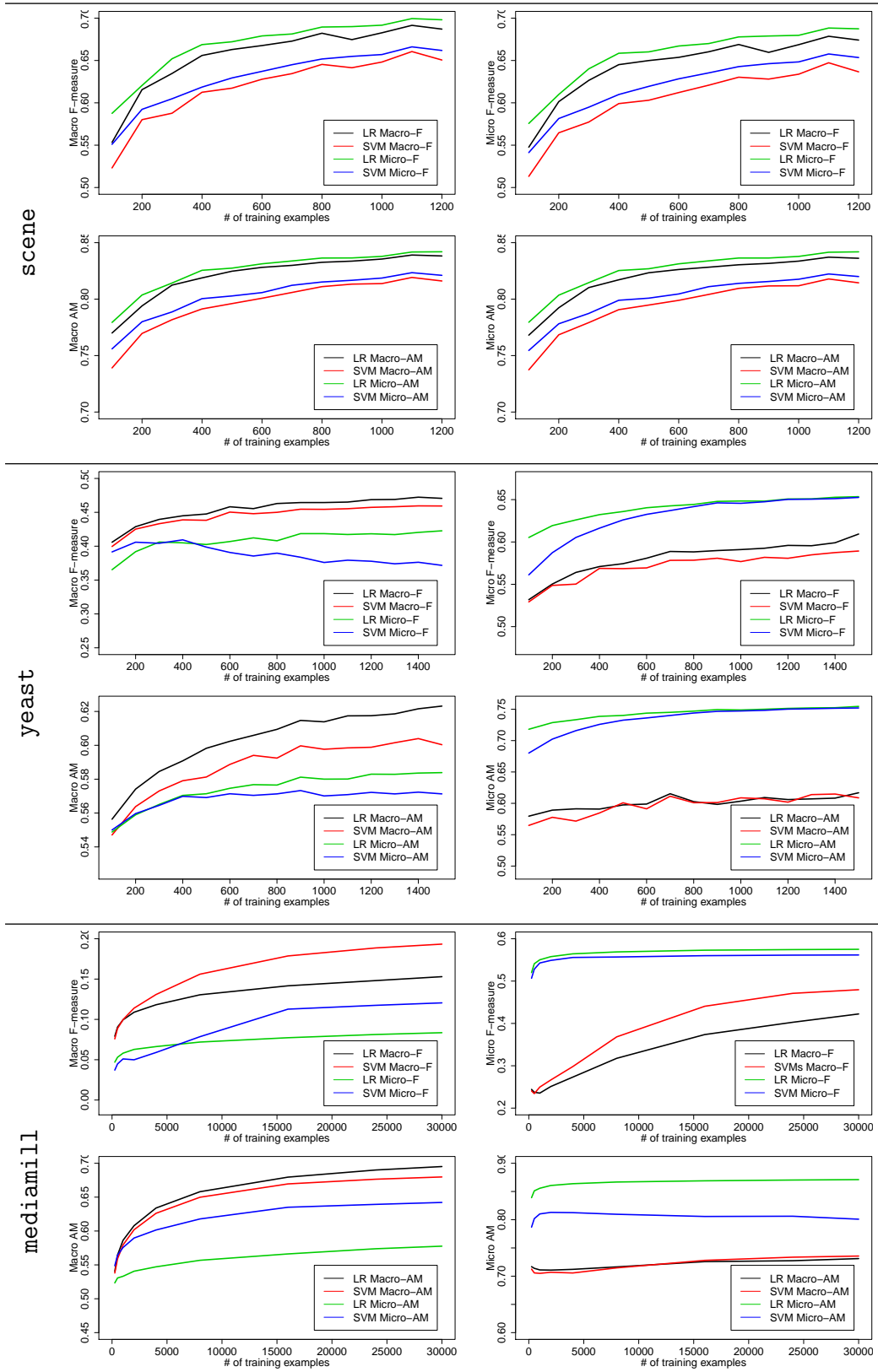


Figure 4: Average test set performance on benchmark data sets for multi-label classification as a function of the number of training examples. Macro- and micro-averaged F-measure and AM are plotted for LR and SVM tuned for all the measures.

## 7 Summary

We present a theoretical analysis of a two-step approach to optimize classification performance metrics, which first learns a real-valued function  $f$  on a training sample by minimizing a surrogate loss, and then tunes the threshold on  $f$  by optimizing the target performance metric on a separate validation sample. We show that if the metric is a linear-fractional function, and the surrogate loss is strongly proper composite, then the regret of the resulting classifier (obtained from thresholding real-valued  $f$ ) measured with respect to the target metric is upperbounded by the regret of  $f$  measured with respect to the surrogate loss. The proof of our result goes by an intermediate bound of the regret with respect to the target measure by a cost-sensitive classification regret. As a byproduct, we get a bound on the cost-sensitive classification regret by a surrogate regret of a real-valued function which holds simultaneously for all misclassification costs. We also extend our results to cover multilabel classification and provide regret bounds for micro- and macro-averaging measures. Our findings are backed up in a computational study on both synthetic and real data sets.

## References

- Agarwal S (2014) Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research* 15:1653–1674
- Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156
- Dembczyński K, Cheng W, Hüllermeier E (2010) Bayes optimal multilabel classification via probabilistic classifier chains. In: *ICML 2010*, Omnipress, pp 279–286
- Dembczyński K, Waegeman W, Cheng W, Hüllermeier E (2012) On loss minimization and label dependence in multi-label classification. *Machine Learning* 88:5–45
- Dembczyński K, Jachnik A, Kotłowski W, Waegeman W, Hüllermeier E (2013) Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In: *ICML*
- Devroye L, Györfi L, Lugosi G (1996) *A Probabilistic Theory of Pattern Recognition*, 1st edn. Springer
- Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874
- Gao W, Zhou ZH (2013) On the consistency of multi-label learning. *Artificial Intelligence* 199-200:22–44
- Hastie T, Tibshirani R, Friedman JH (2009) *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer
- Jansche M (2005) Maximum expected F-measure training of logistic regression models. In: *HLT/EMNLP 2005*, pp 736–743
- Jansche M (2007) A maximum expected utility framework for binary sequence labeling. In: *ACL 2007*, pp 736–743
- Koyejo O, Natarajan N, Ravikumar PK, Dhillon IS (2014) Consistent binary classification with generalized performance metrics. In: *Neural Information Processing Systems (NIPS)*
- Koyejo O, Natarajan N, Ravikumar P, Dhillon IS (2015) Consistent multilabel classification. In: *Neural Information Processing Systems (NIPS)*
- Lewis D (1995) Evaluating and optimizing autonomous text classification systems. In: *SIGIR 1995*, pp 246–254

- Manning CD, Raghavan P, Schütze H (2008) Introduction to Information Retrieval. Cambridge University Press
- Menon AK, Narasimhan H, Agarwal S, Chawla S (2013) On the statistical consistency of algorithms for binary classification under class imbalance. In: International Conference on Machine Learning (ICML)
- Musicant DR, Kumar V, Ozgur A (2003) Optimizing f-measure with support vector machines. In: FLAIRS Conference, pp 356–360
- Nan Y, Chai KMA, Lee WS, Chieu HL (2012) Optimizing F-measure: A tale of two approaches. In: International Conference on Machine Learning (ICML)
- Narasimhan H, Vaish R, Agarwal S (2014) On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In: Neural Information Processing Systems (NIPS)
- Narasimhan H, Ramaswamy HG, Saha A, Agarwal S (2015) Consistent multiclass algorithms for complex performance measures. In: International Conference on Machine Learning (ICML)
- Parambath SP, Usunier N, Grandvalet Y (2014) Optimizing F-measures by cost-sensitive classification. In: Neural Information Processing Systems (NIPS)
- Petterson J, Caetano TS (2010) Reverse multi-label learning. In: Advances in Neural Information Processing Systems 24, pp 1912–1920
- Petterson J, Caetano TS (2011) Submodular multi-label learning. In: Advances in Neural Information Processing Systems 24, pp 1512–1520
- Reid MD, Williamson RC (2010) Composite binary losses. *Journal of Machine Learning Research* 11:2387–2422
- Reid MD, Williamson RC (2011) Information, divergence and risk for binary experiments. *Journal of Machine Learning Research* 12:731–817
- Tsochantaridis I, Joachims T, Hofmann T, Altun Y (2005) Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6:1453–1484
- Waegeman W, Dembczyński K, Jachnik A, Cheng W, Hüllermeier E (2013) On the Bayes-optimality of F-measure maximizers. *Journal of Machine Learning Research* 15:3513–3568
- Zhao MJ, Edakunni N, Pocock A, Brown G (2013) Beyond Fano’s inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research* 14:1033–1090